himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

Himalayan Linguistics

Tibetan functional chunk recognition using statistical based methods

Lin Li

Qinghai Normal University

Congjun Long

Chinese Academy of Social Sciences

Weina Zhao

Qinghai Normal University

ABSTRACT

Functional chunks can not only reveal the structure of a sentence and the relation among chunks but also play an important role in understanding the meaning of a sentence. Therefore recognizing functional chunks is a sub-field of Natural Language Processing, which can effectively improve the performance of syntactic parsing. This paper proposes a Tibetan functional chunk classification that provides the foundation for functional chunk recognition. To test the feasibility of the proposed theory, we observe the distribution of Tibetan functional chunks in our corpus. Statistics reveals that our classification is satisfactory; it is able to describe sentence structure comprehensively. Then we establish a functional chunking model based on Conditional Random Fields (CRFs). After selecting appropriate features, a couple of experiments have been conducted. The *F1* score achieves 82.30 by employing extended features.

KEYWORDS

Tibetan, NLP, functional chunking

This is a contribution from *Himalayan Linguistics, Vol. 15(1): 68–77.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

Tibetan functional chunk recognition using statistical based methods

Lin Li
Qinghai Normal University

Congjun Long
Chinese Academy of Social Sciences

Weina Zhao
Qinghai Normal University

1 Introduction

Syntactic parsing has been a goal of Natural Language Process (NLP) for a long time. To reduce the difficulty of syntactic parsing, chunking (shallow parsing) is widely applied in various NLP systems such as information retrieval and machine translation. Early on Abney (Abney 1991) designed and implemented an English chunk recognition system, and proposed a chunk classification as well. CoNLL-2000 set chunking as a shared task, which has significantly promoted research on chunking (Sang 2000). To acquire more useful information, functional chunk recognition has been proposed (Drabek & Zhou 2001).

Functional chunk recognition aims to reveal sentence structure and acquire semantic information by recognizing the main functional chunks in a sentence. In the past decade, functional chunking has attracted much interest. Most previous work adopts machine-learning methods including HMMs (hidden Markov models) (Freitag & MacCallum 2000), CRFs (conditioned random fields) (Sha & Pereira 2003; Li 2000; Li et al. 2013; Kang et al. 2015; Wang et al. 2015), SVM (support vector machines) (Bie et al. 2008). To improve the performance of statistical methods, increasing linguistic features are applied in chunk recognition (Yao et al. 2007). However, the study of Tibetan functional chunking is not as prosperous as English or Chinese.

In this work, we defined a Tibetan functional chunk classification in the view of computational linguistics, and propose a CRFs based model, which can identify functional chunks effectively. The overall *F1* score reaches 82.30.

2 Tibetan functional chunking

2.1 Classification of Tibetan functional chunks

Although there is a well-established categorization of English chunks as a CoNLL-2000 shared task, the functional chunks discussed in the present work are different. We distinguish 'chunks'

per se from 'functional chunks'. Without the consideration of a larger context, a word sequence may be recognized as a 'chunk' on the basis of its internal structure. In other words, chunks are defined by a bottom-up method; however, 'functional chunks' are defined by a top-down method. A word sequence qualifies as a functional chunk on the basis of its position in context, with its category determined by grammatical relations. Functional chunks are normally longer than chunks and their structure is normally more complex.

We focus on building up a functional chunk classification that is beneficial to further Tibetan functional chunking. We believe that a functional chunk should satisfy two conditions: (1) finiteness, each word of a sentence should be classified into a functional chunk; (2) linearity, each functional chunk of a sentence should be non-overlapping. According to these two principles, we define 7 types of Tibetan functional chunks (Li et al. 2013). A Tibetan corpus is annotated according to our classification, and the distribution of functional chunks is shown in Table1. Furthermore, the annotated corpus will be used to test our functional chunk recognition method later in this paper. The corpus consists of Tibetan simple sentences chosen from textbooks, technical books, and novels.

Functional Chunk	Label	Count	Percentage
Subject	S	3934	18.32%
Predicate	P	5336	24.86%
Object	O	2836	13.21%
Adverbial	D	2689	12.53%
Complement	C	908	4.23%
Syntax Marker	M	4745	22.10%
Independent	I	1020	4.75%
Total		21468	100%

Table1. Distribution of Tibetan functional chunk in our corpus

Because all the following work in this paper is based on this classification, we discuss the distribution of functional chunks a bit more. Our corpus consists of 4,611 sentences (39,824 words) and the distribution of functional chunks in this corpus shown in table1. These statistics show that our classification is able to cover most of the functional chunks found in Tibetan text, and their distribution is as follow: (1) subject, predicate, and syntax marker account to as much as 65%; (2) the proportion of object and adverbial is 25%; (3) the proportion of complement and independent element is relatively small, which is less than 9%. For instance, the following Tibetan sentence is annotated by our classification. In this work, we adopt the Tibetan POS and TAG system developed by Long and Liu (Long & Liu 2016). For instance:

Tibetan:[ཁོང་/rh]{S}[ཕྱི་/wa]{M}[འདི་རྣམ་/rd]{D}[བྱས་/vo]{P}[ན་/c]{X}[རང་/rhཀྱི་/wgཀུན་བདག་/ng]{S}[རྡེབ་བྱ་/a]{C}[ཨ་/ub]{M}[ཀཆོང་/voཞིད་/va]{P}[།/xp][1]

[1] In the example sentence, meaning of POS tag is as follow: *rh* is pronoun, *rd* is adverb, *vo* is verb, *c* is conjunction, *ng* is noun, *a* is adjective, *ub* is auxiliary.

English: [you]{S}[like this]{D}[do]{P}[if]{X}[your situation]{S}[worse]{C}[may get]{P}

If you do something like this, your situation may get worse.

2.2 Tibetan functional chunking annotation system

We take functional chunking as a sequence labeling task, therefore we adopt an improved start/end label collection named T, T={B-S, I-S, E-S, B-O, I-O, E-O, B-D, I-D, E-D, B-C, I-C, E-C, B-P, I-P, E-P, B-M, I-M, E-M, B-X, I-X, E-X, XP}. The first part of a tag means the location of a word, and the second part is the functional chunk type of this word. For instance, B-S means that this word is the first word of the subject chunk.

Tibetan: [ཁྱེད་རང་/rh གི་/wg ཁང་པ་/ng དེ་/rd]{S}[ཨོག་མ་/ng གཤམ་པ་/m ཨི་/wg གཤམ་གྱང་ཚོར་གཞི་/m]{O}[རེད་/vl]{P}[།/xp]

English: [Your room number]{S}[302]{O}[is]{P}

Your room number is 302.

For the convenience of functional chunking, a sentence can be labeled as shown in Table 2.

No.	Word	POS	Label
1	ཁྱེད་རང་	rh	B-S
2	གི་	wg	I-S
3	ཁང་པ་	ng	I-S
4	དེ་	rd	E-S
5	ཨོག་མ་	ng	B-O
6	གཤམ་པ།	m	I-O
7	ཨི་	wg	I-O
8	གཤམ་གྱང་ཚོར་གཞི་	m	E-O
9	རེད་	vl	B-P
10	།	xp	XP

Table2. Tibetan functional chunking sample

3 Tibetan functional chunking model

3.1 Statistical based functional chunking model

Conditional Random Fields (CRFs) (Lafferty 2001) are a class of conditional probability models, which can overcome the labeling bias problem of MEMM (maximum-entropy Markov models). CRFs have been widely used in Natural Language Processing tasks such as word segmentation and part of speech (POS) tagging, named entity recognition, chunking, etc. A number of practices have proved that CRFs are an excellent choice for sequence-tagging (Blaheta 2004), therefore we build up a functional chunking model using CRFs.

CRFs define conditional probability distributions $p(Y|X)$ of label sequences given input sequences. Given an input sequence $X=x_1x_2x_3\ldots x_n$ (x_i presents a word and its POS tag) and a label sequence $Y=y_1y_2y_3\ldots y_n$ ($y_i \in T$, T is the set of functional chunk tags), a CRF of (X,Y) is specified by a local feature vector f and a corresponding weight vector λ. Each local feature is composed of a state

feature $s(y,x,i)$ or an edge feature $t(y_{j-1},y_j,x,i)$, y_{j-1} and y_j are labels, x is an input sequence, and i is an input position. Hence, a CRF defines the conditional distribution as follow:

$$p\_\lambda\ (y \mid x) = \exp(\lambda \cdot F(Y,X)\ )/(Z\_\lambda\ (X)\ )$$

where

$$Z_\lambda(x) = \sum_y (\lambda \cdot F(y,x))$$

$$F(Y,X) = \sum_{i=1}^{n} f_j(y_{j-1}, y_i, x, i)$$

$F(Y,X)$ is a global feature vector, f is a local feature vector, and $Z(x)$ is a normalization factor. As a result, recognition of all functional chunk of X is a process of seeking for the optimal output sequence Y, which can be expressed as the following equation:

$$\hat{y} = argmax p_\lambda(y|x) = argmax \lambda F(y,x)$$

3.2 Features

Theoretically, CRFs could achieve better performance with plentiful features. However, experiments show that too many features could not only increase the complexity but also prolong the training and testing process of CRFs (Sha & Pereira 2003). Therefore it is an important step of functional chunking to build up an optimal feature template. We classify the features that could be applied in functional chunking into two categories: conventional features and extended features.

3.2.1 Conventional features

The conventional feature set contains lexical words and their POS tag only. w_i donates the input token and p_i is its POS tag, i is the relative position, for instance, w_0 is the current token, w_{-1} is the left word of w_o and w_1 is the right word of w_o.

3.2.2 Extended features

We introduce three types of extended features into our chunking model. They are number of syllables, predicate verbs, and word position (Huang & Zhao 2006). The feature of number of syllables can provide the chunking model with rich information about functional chunk internal structure. Also, the predicate verb of a sentence can comprehensively reflect the sentence structure and grammatically framework in Tibetan. Jiang (Jiang 2005) proposes that Tibetan verbs can be divided into 12 categories and establishes correspondence rules between predicate verb and sentence structure. Therefore, we provide the chunking model with the predicate verb as information.

Word position originally refers to the relative position of a character that forms a word in Chinese. Instead, here word position means the relative position of a word in a functional chunk. The feature of word position is a quantitative feature that contributes to identifying functional chunk boundaries.

We represent word position with the annotation set $T2=\{B, I, E, U\}$, B denotes the first word of a functional chunk, E donates the last word, I is a word internal to the functional chunk, and U refers to the one unique word constituting a functional chunk by itself. The word position $t_j \in T2$ of word w_i is defined as follow.

$$P_{w_i}(t_j) = \frac{count(w_i, t_j)}{\sum_{t_j \in T2}(w_i, t_j)}$$

$count(w_i, t_j)$ denotes how many times the word w_i occurs at the position t_j. We specify that if $P_{w_i}(t_j) > 0.7$, then t_j is its main word position; otherwise we consider w_i to not have a main word position. We count $P_{w_i}(t_j)$ of each word in our corpus and the result is shown in Table3.

Word Position	Amount	Percentage
B	911	22.05%
I	311	7.53%
E	538	13.02%
U	1444	34.96%
Total	3204	77.56%

Table3. Word position distribution

The feature of word position can provide the model with effective information, due to the statistics show that as much as 78% words have a main word position. An annotation set *T3*[2] is adopted in our work to provide our chunking model with word position information.

Take [ཉེན་མཚན་/ngཀྱི་/wgདྲོད་ཚད་/ng]{S}[ཁྱད་པར་/ngཆེན་པོ་/a][འདུག/ve] as an example to illustrate our features. When we apply both conventional and extended features into functional chunking, the annotation result is shown in Table4.

Word	POS	NOS	WP1	WP2	WP3	WP4	PV	TOPV	Label
ཉེན་མཚན་	ng	2	BL	IL	EL	UH	འདུག	ve	B-S
ཀྱི་	wg	1	BL	IH	EL	UL	འདུག	ve	I-S
དྲོད་ཚད་	ng	2	TT	TT	TT	TT	འདུག	ve	E-S
ཁྱད་པར་	ng	2	BL	IL	EL	UM	འདུག	ve	B-C
ཆེན་པོ་	a	2	BL	IL	EM	UL	འདུག	ve	E-C
འདུག	ve	1	BL	IL	EM	UL	འདུག	ve	B-P

Table4. Functional chunking annotation Sample

In table4, NOS means 'number of syllables', WP means 'word position', PV means 'predicate verb', and TOPV means 'type of predicate verb'.

[2] A tag of *T3* contains two parts: one is the word position, and the other is the probability of the word position. For instance, *BH* means a word has a high probability (100%, 70%) applied at the beginning of a functional chunk, *BM* means a medium probability (69%, 50%) within a functional chunk, and BL means a low probability (49%, 0%) locating at the ends of a chunk. And if a word does not have a main word position it will be labeled as *TT*.

4 Experiment and results

4.1 Data and evaluation measures

The CRF++ package[3] allows us to apply a large number of features in functional chunking. We manually build up a Tibetan corpus based on the functional chunk classification discussed in Section 2. We conduct a couple of functional chunking experiments using these language resources. Our corpus contains 4611 sentences (39824 words) that can be divided into 21468 functional chunks. In the experiments, we randomly choose 4000 sentences as training data from the corpus, and we use the remaining 611 sentences as a test set. As the CoNLL-2000 shared task did, we evaluate the performance of the Tibetan functional chunking model we proposed using precision (*P*), recall (*R*), and *F1* measure. These evaluation measures are defined as follow:

$$P = \frac{Chunk_{correct}}{Chunk_{return}} \times 100\%$$

$$R = \frac{Chunk_{correct}}{Chunk_{all}} \times 100\%$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

$Chunk_{all}$ is the total number of functional chunks in the test set, $Chunk_{correct}$ is the number of functional chunks correctly recognized by our model, and $Chunk_{return}$ is the number of functional chunks recognized by our model.

4.2 Results

4.2.1 Baseline experiment

We take the model based on conventional features as a baseline in our work. The result of the baseline experiment offers an evaluation criterion for further experiments. The result of the baseline experiment is shown in Table5.

Functional Chunk	*P*	*R*	*F1*	Functional Chunk	*P*	*R*	*F1*
C	70.77%	66.19%	68.40%	D	72.78%	78.95%	75.74%
M	94.50%	95.72%	95.11%	O	61.95%	78.01%	69.06%
P	82.50%	87.82%	85.08%	S	71.25%	83.65%	76.95%
I	68.42%	48.45%	56.73%	Overall	74.60%	76.97%	75.30%

Table5. Result of baseline experiment

[3] http://cefpp.sourceforge.net

The overall *F1* on test set is 75.30%, which proves that our choice of CRFs to identify Tibetan functional is practical. The result shows that the *F1* of syntax marker (*M*) chunk substantially exceeds the overall *F1*. The reason why our model recognizes *M* well is that the structure of *M* is simple and Tibetan syntax markers form a closed set. The model performed worst in recognizing independent (*I*) chunks, because of their low frequency and the structural complexity of complements in our corpus.

4.2.2 Combined conventional feature experiment

We only employ word and POS as unigram features in the baseline experiment. We propose Template A that adds bigram features and combined features into the Template Baseline in this experiment. The bigram features bring state transition information into our model. The combined feature refers to the feature composed of word and POS together. We believe that the combined feature can provide our model with rich information.

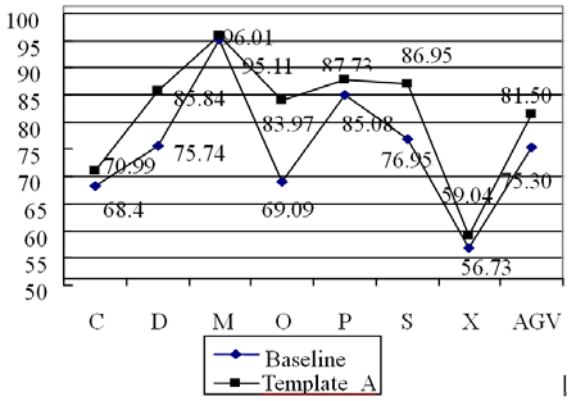

Figure1. Result of baseline experiment and combined feature experiment

According to the result in Figure1, we find that the overall *F1* of Combined conventional feature experiment has been improved to 81.50%. Yet, combined features bring too many features into our model, which leads to the experiment becoming time-consuming. Therefore, we only keep the bigram feature in the remaining experiments, leaving aside the combined feature.

4.2.3 Extended feature experiment

We incorporate each extended feature to our model separately, and observe its effects on the model performance. The results are shown in Table6.

Extended feature	P	R	F1
Baseline	74.60%	76.97%	75.30
Number of Syllable	80.47%	80.47%	80.47
Word Position	83.91%	81.58%	82.53
Predicate Verb	84.31%	80.91%	82.30

Table6. Results of extended feature experiment

By applying the extended features to the model, the model has achieved varying degrees of improvement. The *F1* has been improved 6% on average; the results are listed in Table7 in detail.

Functional Chunk	Baseline	Simple Features	Word Position	Number of Syllable	Predicate Verb
C	68.4	70.99	70.77	73.59	72.31
D	75.74	85.84	89.16	85.92	85.26
M	95.11	96.01	95.09	95.64	95.74
O	69.06	83.97	83.61	84.58	87.65
P	85.08	87.73	88.51	87.76	87.83
S	76.95	86.95	87.37	87.79	88.475
I	56.73	59.04	63.22	59.6	58.82
Total	75.3	81.5	82.53	80.47	82.30

Table7. Results of experiments adopting extended features

4.3 Analysis of results

Functional chunking based on CRFs has advantages in recognition speed and precision, yet it still commits some mistakes. In this section, we summarize and analyze these mistakes, which can benefit future research. Almost all mistakes fall into the following categories:

4.3.1 Subject and object boundary and type recognition mistakes

The basic word order of Tibetan is subject, object and verb. In most cases, the subject is immediately adjacent to the object in a sentence. Furthermore, normally there is no obvious syntax mark between subject and object.

Our recognition result:[མར་/ng རྫོག་/q འདི་/rd རྒྱ་མ་/q གསུམ་/m]{S}[ཏག་ཏག་/d]{D}[རེད་/vl]{P}

Standard recognition result: [མར་/ngརྫོག་/q འདི་/rd]{S}[རྒྱ་མ་/q གསུམ་/m ཏག་ཏག་/d]{O}[རེད་/vl]{P}
English:Have just 10 jin ghee.

The recognition result is a typical type of mistake of our model. The model does not distinguish the object from the subject. We speculate that this kind of mistake can be attributed to two causes: (1) the similarity of subject and object structures; (2) the lack of an obvious mark between these two functional chunks.

4.3.2 Complement chunk recognition mistakes

The proportion of complement chunks is comparatively small in our corpus. Moreover, the constituents of the complement chunks are similar to adverbial chunks. As a consequence, the model is inclined to identify complement chunks as adverbial chunks, like the result of example.2.

Our recognition result: [རྒྱ་ས/ns ཏེ་/wg བོད་ཟས་/ng]{S}[ཏུ་ཙང་/d ཞིམ་མོ་/a]{C}[འདུག/ve]{P}

Standard recognition result: [རྒྱ་ས/ns ཏེ་/wg བོད་ཟས་/ng]{S}[ཏུ་ཙང་/d]{D}[ཞིམ་མོ་/a]{C}[འདུག/ve]{P}

English: The Tibetan food in Lhasa is very yummy.

5 Conclusions

One solution for Tibetan functional chunking is presented in this paper, which mainly includes the following: (1) proposal for a Tibetan functional chunk classification; (2) establishing a functional chunking based on CRFs; (3) explore and introduce effective extended information into the model. We conduct a couple of experiments, and acquire an overall *F1* of 82.30.

Comparing with similar work in Chinese or English, the recognition performance shown in this paper is still weak. On the one hand, the disadvantage is caused by the limitation of corpus size. Therefore, we plan to accumulate more Tibetan text with annotation in the future. On the other hand, introducing more useful features could be helpful to improve the model.

Acknowledgement

This research was supported by the Qinghai Natural Science Foundation under Grant 2015-ZJ-923Q，Ministry of Education under Grant Z2015066 and Z2015067, and National Natural Science Foundation of China under Grant 61550004.

References

Abney, Steven. 1991. "Parsing by Chunks". In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-based parsing*. Norwell, MA, USA: Kluwer Academic Publishers, 257-278.

Blaheta D. 2004. *Function tagging*. Brown University, PhD dissertation.

Drabek, E. F., and Q. Zhou. 2001. "Experiments in learning models for functional chunking of Chinese text". *Systems, Man, and Cybernetics, IEEE International Conference on IEEE* (2): 859-864.

Freitag, Dayne, and A. McCallum. 2000. "Information extraction with HMM structures learned by Stochastic Optimization". In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*. Austin, Texas.

Huang, Changning, Hai Zhao. 2006. "New method of Chinese word segmentation by word formation". *Advanced development of Chinese information processing*, Beijing: Tsinghua University Press. (黄昌宁,赵海. 2006. 由字构词——中文分词新方法[A].中文信息处理前沿进展[C], 北京:清华大学出版社.)

Jiang, Jiang. 2005. "The classification of Tibetan verbs and relative patterns based on semantics and syntax". *Journal of Chinese Information Processing* 20.1: 37-43.(江荻. 现代藏语动词的句法语义分类及相关语法句式[J]. 中文信息学报, 2006, 20(1):37-43.)

Kang, Caijun, Congjun Long, Di Jiang. 2015. "Tibetan names recognition research based on CRF". *Computer Engineering and Application* 3: 109-111.(康才畯, 龙从军, 江荻. 2015. 条件随机场的藏文人名识别研究[J]. 计算机工程与应用, (3):109-111.)

Lafferty, J., McCallum A, Pereira F C N. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". *Proceedings of the 18th International Conference on*

Machine Learning 2001 (ICML 2001). San Francisco: Morgan Kaufmann Publishers Inc.: 282-289

Long, Congjun, Huidan Liu. 2016. *Study on the theory and practice of Tibetan word segmentation*. Beijing: Intellectual property press. (龙从军，刘汇丹.2016.藏文自动分词的理论与实践研究. 知识产权出版社)

Li, Ru, Lijun Zhong, Shuanghong Li, Zezheng Zhang. 2010. "Automatic identification of Chinese multiword chunk based on CRF". *Web Intelligence and Intelligent Agent Technology* (WI-IAT), 2010 IEEE/WIC/ACM International Conference on IEEE: 174-177.

Li, Lin, Congjun Long, Di Jiang. 2013. "Tibetan functional chunks boundary detection". *Journal of Chinese Information Processing* 27.6: 165-168.(李琳, 龙从军, 江荻. 2013. 句法功能组块的边界识别[J]. 中文信息学报, 27(6):165-168.)

Sang, T. K. 2000. "Introduction to the CoNLL-2000 shared task: Chunking". *Proceedings of CoNLL-2000 and LLL-2000 Conference*. Lisbon, Portugal: 127-132.

Sha, F, Pereira F. 2003. "Shallow parsing with conditional random fields". *Proceedings of Human Language Technology. North American Chapter of the Association for Computational Linguistics Annual Meeting*. Edmonton: 213-220.

Wang, Tianhang, Shumin Shi, Congjun Long, Heyan Huang, and Lin Li 2015. "Tibetan chunking based on error-driven learning strategy". *Journal of Chinese Information Processing* 5: 170-175.(王天航, 史树敏, 龙从军,等. 2014. 基于错误驱动学习策略的藏语句法功能组块边界识别[J]. 中文信息学报, (05):170-175.)

Yao, Limin, M. Li, and C. Huang. 2007. "Improving Chinese chunking with enriched statistical and morphological knowledge". *International Conference on Natural Language Processing & Knowledge Engineering*, Beijing: 149-156.

Zhi Bie, Junsheng Zhou, Jiajun Chen. 2008. "SVM-Adaboost based Chinese text chunking". *Computer Engineering and Applications* 44.21:171-173. (别致, 周俊生, 陈家骏. 基于 SVM-Adaboost 的中文组块分析[J]. 计算机工程与应用, 2008, 44(21):171-173.)

Weina Zhao
zhaoweina1999@qq.com